Similar to the research methods we examined in Chapter 10, the data used for the methods described in this chapter often rely on secondary data. Importantly, the methods covered in this chapter can be used both for investigative purposes and for basic research purposes. The rise of social media platforms such as Facebook and Twitter have provided both researchers and practitioners with an abundance of data from which to draw inferences.

In this chapter, we will first introduce you to social network analysis and demonstrate how it has been applied to research as well as law enforcement investigative practices. We next discuss mapping generally (and crime mapping specifically) and examine how this technique has also been applied to police practice as well as general research. And finally, we will conclude the chapter with an introduction to the concept of big data, which you will learn is not merely an extremely large and dynamic set of data but also refers to techniques used to extract information from these datasets.

**SOCIAL NETWORK ANALYSIS**

It is virtually impossible today not to be part of social networks. Everyone you interact with, including those in the virtual as well as in the real world, are part of your social network. We inherently think about the world in terms of these networks, including such networks as familial and friendship networks, other students in your major in college, people who work out at the same time as you at the gym, and the many “friends” you may not actually know in real life on platforms such as Facebook or those who similarly liked something on Twitter. The method of *social network analysis* (SNA) has increasingly been used since the Internet and these social media platforms emerged. There are entire textbooks devoted to SNA, so the goal of this chapter is to introduce you to the basics, along with a few case studies that highlight their applicability in the field. SNA is not one type of method but is an approach to analysis and a set of methodological techniques that help researchers describe and explore relationships that both individuals and groups have with each other (Scott 2017).

*Social networks* are types of relationships that can include many different forms, such as face-to-face and online interactions, digital economic transactions, interaction with a criminal justice agency, geopolitical relations among nation states, and so on. As you can see, there are numerous types...
of networks available for analysis. The most important component of any network is that it is relational. That is an important assumption. So far in this book, we have talked about variables that measure attributes of the units of analysis, such as the behavior of people, the crime rates of cities, and so on. **Relational data** measure the contacts, connections, attachments, and ties that relate one unit to the next (Scott 2017). As a result, these relational data are not properties of any particular unit (e.g., individual, group, city) but are “relational systems” of units that are created by connecting pairs of interacting units. Importantly, then, it is the technique used to describe and examine relational data that is the key to SNA. Many of the traditional methods that we have already discussed in this book can also be used to collect relational data. For example, surveys, interviews, participant observation, and secondary data can all be used to generate relational data for SNA, as we will see in the case studies that follow.

Although literally thousands of articles examined aspects of social structure in the early twentieth century, one of the first graphical applications of social networks was created by Jacob Moreno to examine friendship choices. In his classic book, *Who Shall Survive?* Moreno (1953) describes his definition of *sociometry* as being in accordance with its etymology from Latin and Greek, “with the emphasis . . . on the second half of the term, ‘metrum,’ meaning measure, but also on the first half, ‘socius’ meaning companion” (31). Instead of focusing exclusively on the individual or exclusively on an aggregate entity, Moreno believed that the relationship between individuals within a group must also be examined. Using a sociometric test, which required individuals to choose their associates from a group in which they were a member, *attractions* and *repulsions* were determined. For example, if your instructor wanted to understand the social structure of the class you are in right now, she or he might ask each student to hypothetically choose among the students whom they wanted to have sit next to them (attraction) and whom they would like to have moved to another class (repulsion). Responses received from each individual in the group could then be graphed in a **sociogram**, which is a way of representing social configurations, with individuals (or some other unit) represented by points and their social relationships to one another depicted by lines (Moreno 1953). The formal terminology that describes these graphs as well as the units and relationships therein are called several things, depending on the discipline. The social sciences generally call the basic units in a graph **nodes** (sometimes called *actors* or *vertices*) and nodes are connected by **relations** (sometimes called *ties*, *links*, *arcs*, or *edges*). As noted above, relationship data such as these can be collected from many places, such as official records, Facebook friends, and so on (Yang, Keller, and Zheng 2017).

SNA usually consists of at least two datasets. The first is called the **nodelist**, where all of the units of observation are stored. The second defines the relations between these units. One of the most common types of relations data is called an **adjacency matrix** (sometimes called *network matrix*), wherein the nodes constitute both the rows and the columns and the cells specify if and what kind of relationship exists between the nodes at the intersection of each row and column. We are going to stick with the simplest case of a **binary network**, which only distinguishes whether a relation does or does not exist between a pair of nodes (Yang et al. 2017). An example of a network graph and the nodelist and adjacency matrix upon which it is based is presented in Exhibit 11.1. As you can imagine, nodes and relations can be much more complicated than this simple example, and special software is required to mathematically describe the numerous networks that emerge from such data. A discussion of these issues is beyond the scope of this text, but we want to provide you with some exciting case studies of how SNA is being used in research related to criminology and criminal justice.
CASE STUDY

Networks of Terrorist Cells

On September 11, 2001, nineteen members of the Islamic extremist group al-Qaeda hijacked four airplanes and carried out suicide crashes into four places in the United States. The targets included the north and south towers of the World Trade Center, the Pentagon, and Washington, D.C. (in the last instance, the passengers fought the hijackers, preventing the plane from hitting its target and causing it to crash into an empty field in Pennsylvania instead). Almost 3,000 people were killed, including all of the passengers on the planes along with the 19 hijackers and many hundreds of people in the targeted buildings, which included rescuers. About a month before the 9/11 attack, Zacarias Moussaoui, a French citizen of Moroccan descent, and sometimes referred to as the 20th hijacker, was arrested after he raised suspicion at a flight school in Oklahoma by requesting information on flying a 747. Moussaoui was eventually indicted and found guilty in 2006 of six charges, including conspiracy to commit acts of terrorism. Information from his indictment (United States of America v. Zacarias...
Moussaoui 2001), along with other information uncovered by The New York Times and the Washington Post have been used to conduct social network analyses of the terrorist cell that carried out 9/11. One of the first attempts was made by Krebs (2002), who created one of the first network graphs of the terrorist network. An adapted snippet of his graph is displayed in Exhibit 11.2.

Without going into the advanced statistical analysis performed by Krebs to describe the strength of the relations for the terrorists, many conclusions can be drawn from this graph, including the meticulousness with which the hijackers kept their identities unknown even from each other. Krebs explains, “Many pairs of team members were beyond the horizon of observability. . . . Keeping cell members distant from each other, and from other cells,
minimizes damage to the network if a cell member is captured or otherwise compromised” (2002, 46). Kreb’s graph also confirms the fact that Mohamed Atta was the likely leader of the cell, as he has the most relations with other nodes.

Krebs analysis shows the benefits of SNA when putting together a case for prosecution. It also highlights the inherent difficulty of using SNA for preventing or uncovering secret illegal networks. Krebs concludes,

The best solution for network disruption may be to discover possible suspects and then, via snowball sampling, map their ego networks—see whom else they lead to, and where they overlap. To find these suspects it appears that the best method is for diverse intelligence agencies to aggregate their information—their individual pieces to the puzzle—into a larger emergent map. (51)

While using SNA for investigative purposes has its challenges, our next case study demonstrates the merits of doing so.

**CASE STUDY**

**Finding a Serial Killer**

The Green River serial killer (GRK) killed his first victim, 16-year-old Wendy Coffield, in Kings County, Washington. Her body was found in July of 1982 in the Green River, which became the name given to the then-unidentified killer. After 48 other murders, Gary Leon Ridgway was finally charged as the serial killer in September 2001, despite the fact that he was on the list of suspects much earlier. Media interest of the cases generated thousands of leads, and these leads compounded with every new victim, which resulted in a huge amount of data to examine. However, large amounts of data are not the only hindrance to solving a case. Like all of us, police detectives can have cognitive biases (see Chapter 1 for a list of common biases), and when combined with an overabundance of information coming from the public, both reliable and unreliable, investigations can go awry.

In a recent paper, Bichler, Lim, and Larin (2017) have demonstrated how SNA can be used to aid in connecting the pieces of a growing body of evidence. Bichler and her colleagues collected data about the Green River murders from multiple sources, including newspaper reports, books, and court transcripts of Ridgway’s trial. They then performed a SNA analysis of the evidence over time to determine if SNA could have prevented investigators from keeping another man on top of their suspect list instead of Gary Ridgway. They state, “We argue that by identifying which actors shift in structural position during the investigation, it may be possible to reduce the damaging effects of tunnel vision, emphasis on specific evidence, and intuition” (141). The goal of their analysis was to find the connections between the victims and the suspects, along with the places they frequented. Because these murders likely involved strangers, Bichler et al. (2017) explain that other sources of information must be added to find clusters. They state,

People sharing social space will emerge when we link witnesses, friends, and associates to the places they frequent, but finding individuals positioned between different clusters will be even more informative. . . . Who else but the suspect in a crime series would link the victims from different social spaces. (141)

The categories of the data Bichler and her colleagues examined included victims, suspects, investigatory involvement (e.g., witnesses, body finders), family and associates, body disposal sites, last seen locations, and other investigative material. Because there were different levels of
geographical aggregation (e.g., a red light district, a specific hotel), all data points were placed within a census track. Without going into the statistical details of the study, they were particularly interested in “middleman” who connected others not directly linked to each other. This is statistically measured with a betweenness centrality score. The researchers also created multiple graphs using data over time, beginning with graphs using data that was available from the beginning of the investigation and creating more graphs until the data were exhausted (about 30 months into the investigation). Exhibit 11.3A displays a graph using data from the first six months of the investigation while Exhibit 11.3B depicts the investigation at 18 months, after which there was not much new data to incorporate. Line thickness indicates the number of shared places between pairs, diamonds depict suspects, gray circles indicate victims, and other circles represent other case nodes (e.g., witnesses, body disposal sites).

Exhibit 11.3  Network Graph of Nodes in the Green River Killer Investigation at 6 and 18 Months

---

Source: Adapted from Bichler, Lim, and Larin (2017), Table 3, p. 146.
The size of the shapes in the graphs (e.g., circles, diamonds) indicate the degree of between-ness centrality. The highest-scoring individual at the beginning of the investigation was Melvyn Foster, who is represented by the square containing the large diamond in the Exhibit 11.3A. By 18 months, however, Gary Leon Ridgway becomes more noticeable (depicted by the shaded triangle in the second box). Unfortunately, the task force did not focus on Ridgway but continued to focus exclusively on Foster. Bichler and her colleagues concluded,

It is hard to know in hindsight if these results would sway the organization momentum that led investigators to focus on Foster to the exclusion of Ridgway. . . . It might have prevented the working hypothesis from solidifying so early on in the investigation by encouraging members of the Green River Killer Task Force to pay greater attention to Gary Leon Ridgway.

While this research cannot predict what would have happened had SNA been available then, it certainly highlights the utility of SNA for investigative purposes today. We move on next to a research technique that is frequently being used for predictive purposes in law enforcement.

**CRIME MAPPING**

Many of us have adopted an image of crime mapping that involves a police precinct wall with pushpins stuck all over it identifying the location of crime incidents. Shows such as *The Wire*, *The District*, and episodes of *CSI* have also presented the drama behind the use of crime mapping in generating crime counts within various police beats. Crime mapping for intelligence-led policing has been increasing in the past few decades, but crime mapping for research purposes has a very long history; it is generally used to identify the spatial distribution of crime along with the social indicators such as poverty and social disorganization that are similarly distributed across areas (e.g., neighborhood, census tract). Boba (2009) defines crime mapping as “the process of using a geographic information system (GIS) to conduct special analysis of crime problems and other police-related issues” (7). She also describes the three main functions of crime mapping:

1. It provides visual and statistical analyses of the spatial nature of crime and other events.
2. It allows the linkage of crime data to other data sources, such as census information on poverty or school information, which allows relationships among variables to be established.
3. It provides maps to visually communicate analysis results.

Although applied crime mapping similar to this has been used for over one hundred years to assist the police in criminal apprehension and crime prevention, the advent of computing technology has enabled crime mapping to become an advanced form of statistical data analysis. The geographic information system (GIS) is the software tool that has made crime mapping increasingly available to researchers since the 1990s.

Today, crime mapping is being used by the majority of urban law enforcement agencies to identify crime hot spots (Caplan and Kennedy 2015). Hot spots are geospatial locations within jurisdictions where crimes are more likely to occur compared to other areas. Being able to understand where crime is more likely to occur helps agencies deploy resources more effectively, especially for crime prevention purposes. These hot spots can be specific addresses, blocks, or even clusters of blocks (Eck et al. 2005). Of course, crime mapping data with insight from criminological theory is the ideal. As Eck and his colleagues explain, “Crime
theories are critical for useful crime mapping because they aid interpretation of data and provide guidance as to what actions are most appropriate” (3). This is important because the ability to understand why crimes are occurring has a great deal to do with underlying factors related to the environment in which they occur. Kennedy, Caplan and Piza (2012) provide a very illuminating example:

A sole analytical focus on crime hotspots is like observing that children frequently play at the same place every day and then calling that place a hotspot for children playing, but without acknowledging the presence of swings, slides, and open fields—features of the place (i.e., suggestive of a playground) that attract children there instead of other locations absent such entertaining features. (245–46)

Through various symbols, maps can communicate a great deal of information. Exhibit 11.4, which was published by the National Institute of Justice, displays some common symbols used by crime analysts (Eck et al. 2005). As the map shows, dots (A) point to specific places where crime is likely to occur, a crime site (B and C) indicates where crime is equally likely to occur within a particular site, and a crime gradient (D) indicates that the probability of crime is most likely inside the site and decreases as you move toward the edge of the site.

Jennifer A. Herbert, MA, Crime Intelligence Analyst, Crime Analysis and Strategic Evaluation Unit

Jennifer Herbert graduated with a double major in political science and justice studies from James Madison University in 2007. She had aspirations of becoming a police officer and eventually a detective. She was hired as a police officer after graduation, but she realized while at the police academy that she wanted to pursue the crime analysis career path in law enforcement. She became a crime analyst at Chesterfield County Police Department in Virginia. While working full time as an analyst, Jennifer pursued a master's degree in intelligence at the American Military University. She then accepted a promotion to crime intelligence analyst at Henrico County Police Division.

After working as a crime analyst for six years, Jennifer cannot imagine doing anything else. Every day is different when working as a crime intelligence analyst. Some days, Herbert analyzes phone records and maps a suspect's whereabouts during the time of a crime. Other days, she maps the latest residential burglary trend and predicts where the next burglary will occur. She also completes research projects that examine quality-of-life issues for the community, including estimating crimes per 1,000 residents by neighborhood. Herbert's role as a crime analyst is equally important in preventing crime and in helping patrol officers to apprehend offenders. She thinks the most rewarding part of her job is helping people who have been victimized by apprehending offenders and improving the quality of life for county residents. Jennifer has some good advice for students interested in careers involving analysis:

If crime analysis interests you, ask your local police department if you can do an internship (paid or unpaid) to gain experience. Be sure to network with other crime analysts and let them know you are interested in pursuing a career in crime analysis. Courses in all forms of data analysis and GIS (geographic information systems) are almost essential to a career in crime analysis. Even if you did not take GIS classes during your undergraduate studies, many community colleges offer introductory and advanced classes in GIS. Other qualifications that will help you stand out as an applicant include competency in basic statistics and proficiency in data analysis programs, including Microsoft Excel, Access, and SPSS.
CASE STUDY

Social Disorganization and the Chicago School

As we noted earlier, crime mapping for general research purposes has a long history in criminological research. Although they were not the first to use crime mapping, Shaw and McKay (1942) conducted a landmark analysis in criminology on juvenile delinquency in Chicago neighborhoods back in the 1930s. These researchers mapped thousands of incidents of juvenile delinquency and analyzed relationships between delinquency and various social conditions such as social disorganization. After analyzing rates of police arrests for delinquency, Shaw and McKay used police records to determine the names and addresses of those arrested in Chicago between 1927 and 1935. They observed a striking pattern that persisted over the years, as shown in Exhibit 11.5.

Exhibit 11.5 displays one of the maps Shaw and McKay (1942) created that illuminates the spatial distribution of delinquency within concentric circles of Chicago that spread out to the suburbs from the city’s center. As noted in Exhibit 11.5, there is a linear decrease in rates of delinquency as the distance from the Loop (city center) increases. When rates of other community characteristics were similarly mapped (e.g., infant mortality, tuberculosis cases, percentage of families who own their own homes, percentage of...
foreign-born residents, percentage of families on relief), the conclusions were obvious. Shaw and McKay concluded,

It may be observed, in the first instance, that the variations in rates of officially recorded delinquents in communities of the city correspond very closely with variations in economic status. The communities with the highest rates of delinquents are occupied by these segments of the population whose position is most disadvantageous in relation to the distribution of economic, social, and cultural values. Of all the communities in the city, these have the fewest facilities for acquiring the economic goods indicative of status and success in our conventional culture. . . . In the low-income areas, where there is the greatest deprivation and frustration, where, in the history of the city, immigrant and migrant groups have brought together the widest variety of divergent cultural traditions and institutions, and where there exists the greatest disparity between the social values to which the people aspire and the availability of facilities for acquiring these values in conventional ways, the development of crime as an organized way of life is most marked. (318–19)

Exhibit 11.5  Zone Rates of Police Arrests in Chicago, 1931

Source: Shaw and McKay 1942.
CASE STUDY

Predicting Break and Entries (BNEs)

Contemporary researchers and law enforcement officials interested in issues related to crime and criminology have access to more sophisticated computer technology that allows for the creation of more enhanced crime maps. In fact, the easy availability of mapping tools, including mobile GIS technology, is providing many more opportunities for intelligence-led policing, which includes using “data, analysis, and criminal theory . . . to guide police allocation and decision making” (Fitterer, Nelson, and Nathoo 2015, 121). The purpose of crime maps, however, remains the same: to illuminate the relationship between categories of crime and corresponding characteristics such as poverty and disorganization across given locations.

Break and entries (BNEs) are one type of crime that are patterned; in fact, one of the things we know about them is that the probability of a repeat BNE increases for either the original BNE site and/or for homes near it for several weeks after the original BNE. To determine the effectiveness of mapping in predicting residential and commercial BNEs, Fitterer and her colleagues (2015) used data from the Vancouver Police Districts (VPD) alongside data on other characteristics of the districts, including variables such as population density, property values, dominant housing types, and street light density.

Using BNE data by location, hour, day, month, and year from 2001 to 2012, the researchers then created a map of the Vancouver area composed of 200m by 200m grids, placing the data within each grid. For example, Exhibit 11.6 displays the map depicting residential BNE hot spots for the time period. The goal of their research, however, was not merely to describe the BNE data but to use data from the early time period to predict later occurrences of BNEs. For example, Fitterer and her colleagues found that the probability of a repeat BNE occurring up to 850m from the originating BNE increased 53% for the next 24 hours after the first event. While there was still an increase in the likelihood of near BNEs over time, this increased risk decreased to 24% after a week of the original BNE. Importantly, the proportion of historical crime did significantly predict future crime. The authors concluded,

We found [that] both residential and commercial crimes had a strong spatial clustering over short time periods, suggesting a near-repeat offense dynamic . . . [indicating] that perpetrators prefer to reoffend where they have local knowledge about residents’ routine activities, possessions, and can confirm successful property entry. (130)

Exhibit 11.6  Vancouver’s 2001–2011 Break and Enter Hot Spot Map for Residential Break-Ins

Source: Adapted from Fitterer et al. (2015), Figure 9, page 129.
COMPSTAT AND THE NEW YORK CITY POLICE DEPARTMENT

The New York City Police Department introduced CompStat in the 1990s to map crime and measure police activities. This story examines how CompStat is now being brought into the mobile age and being incorporated into surveying New York residents about their fear of crime and trust in police. New Yorkers will receive short sets of questions measuring three themes: Do you feel safe in your neighborhood? Do you trust the police? Are you confident in the New York Police Department? A private company will then analyze the data and send the results to police precinct commanders on a scale from 100 to 900, although the story does not provide the details on how these scores are to be interpreted. The city says questions will be sent to devices that are already open to advertisers and responses will come back anonymously.

For Further Thought:

1. Does this new practice raise ethical questions about how these data will be accessed and used?
2. How can the mobile device be validly placed within a particular precinct? What other sampling issues arise after reading this story?


CASE STUDY

Using Google Earth to Track Sexual Offending Recidivism

While the GIS software that was utilized by Fitterer et al. (2015) has many research advantages for displaying the spatial distributions of crime, other researchers have begun to take advantage of other mapping tools, including Google Earth. One such endeavor was conducted by Duwe, Donnay, and Tewksbury (2008), who sought to determine the effects of Minnesota’s residency restriction statute on the recidivism behavior of registered sex offenders. Many states have passed legislation that restricts where sex offenders are allowed to live. These policies are primarily intended to protect children from child molesters by deterring direct contact with schools, day care centers, parks, and so on. Most of these statutes are applied to all sex offenders, regardless of their offending history or perceived risk of reoffense.

The impact of such laws on sexual recidivism, however, remains unclear. Duwe et al. (2008) attempted to fill this gap in our knowledge. They examined 224 sex offenders who had been reincarcerated for a new sex offense between 1990 and 2005 and asked several research questions, including “Where did offenders initially establish contact with their victims, and where did they commit the offense?” and “What were the physical distances between an offender’s residence and both the offense and first contact locations?” (488). The researchers used Google Earth to calculate the distance between an offender’s place of residence, the place where first contact with the victim occurred, and the location of offense.

Duwe and his colleagues (2008) investigated four criteria to classify a reoffense as preventable: (1) the means by which offenders established contact with their victims, (2) the
distance between an offender’s residence and where first contact was established (i.e., 1,000 feet, 2,500 feet, or 1 mile), (3) the type of location where contact was established (e.g., was it a place where children congregated?), and (4) whether the victim was under the age of 18. To be classified as preventable through housing restrictions, an offense had to meet certain criteria. For example, the offender would have had to establish direct contact with a juvenile victim within one mile of his residence at a place where children congregate (e.g., park, school).

Results indicated that the majority of offenders, as in all cases of sexual violence, victimized someone they already knew. Only 35% of the sex offender recidivists established new direct contact with a victim, but these victims were more likely to be adults than children, and the contact usually occurred more than a mile away from the offender’s residence. Of the few offenders who directly established new contact with a juvenile victim within close proximity of their residence, none did so near a school, a park, a playground, or other locations included in residential restriction laws.

The authors concluded that residency restriction laws were not that effective in preventing sexual recidivism among child molesters. Duwe et al. (2008) stated,

Why does residential proximity appear to matter so little with regard to sexual reoffending? Much of it has to do with the patterns of sexual offending in general. . . . Sex offenders are much more likely to victimize someone they know. For example, one of the most common victim–offender relationships in this study [for those who victimized children] was that of a male offender developing a romantic relationship with a woman who had children. . . . They used their relationships with these women to gain access to their victims. . . . It was also common for offenders to gain access to victims through babysitting for an acquaintance or co-worker. (500)

Clearly, the power of mapping technologies has changed not only the way law enforcement officials are preventing crime but also the way in which researchers are examining the factors related to crime and crime control.

**BIG DATA**

When do secondary data become what is now referred to as big data? **Big data** is a somewhat vague term that has been used to describe large and rapidly changing datasets and the analytic techniques used to extract information from them. It generally refers to data involving an entirely different order of magnitude than what we are used to thinking about as large datasets. For our purposes, *big data* is simply defined as a very large dataset (e.g., contains thousands of cases), accessible in a computer-readable form, that is used to reveal patterns, trends, and associations among variables. The technological advancements in computing power over the past decade have made analyses of these huge datasets more available to everyone, including government, corporate, and research entities alike. Importantly, many researchers now contend that “big data holds great promise for improving the efficiency and effectiveness of law enforcement and security intelligence agencies” (Chan and Moses 2017, 299).

Here are some examples of what now qualifies as big data (Mayer-Schönberger and Cukier 2013): Facebook users upload more than 10 million photos every hour and leave a comment or click on a “like” button almost three billion times per day; YouTube users upload more than an hour of video every second; Twitter users are sending more than 400 million tweets per day. If all this and other forms of stored information in the world were printed in books, one estimate in 2013 was that these books would cover the face of the earth 52 layers thick. That’s “big.”

All this information would be of no more importance than the number of grains of sand on the beach except that these numbers describe information produced by people, available to social scientists, and manageable with today’s computers. Already, big data analyses are being used to predict the spread of flu, the behavior of consumers, and the prevalence of crime.
Here’s a quick demonstration: We talked about school shootings in Chapter 1, which are a form of mass murder. We think of mass murder as a relatively recent phenomenon, but you may be surprised to learn that it has been written about for decades. One way to examine inquiries into mass murder is to see how frequently the term *mass murder* has appeared in all the books ever written in the world. It is now possible with the click of a mouse to answer that question, although with two key limitations: We can only examine books written in English (and in a few other languages) and, as of 2014, we are limited to “only” one quarter of all books ever published—a mere 30 million books (Aiden and Michel 2013).

To check this out, go to the Google Ngrams site (https://books.google.com/ngrams), type in *mass murder* and *serial murder*, and check the case-insensitive box (and change the ending year to 2015). Exhibit 11.7 shows the resulting screen (if you don’t obtain a graph, try using a different browser). Note that the height of a graph line represents the percentage that the term represents of all words in books published in each year, so a rising line means greater relative interest in the term, not simply more books being published. You can see that *mass murder* emerges in the early 20th century, while *serial murder* did not begin to appear until much later, in the 1980s. It’s hard to stop checking other ideas by adding in other terms, searching in other languages, or shifting to another topic entirely. Our next case study illuminates how law enforcement agencies are also harnessing big data to make predictions.

Exhibit 11.7 Ngram of Mass Murder and Serial Murder


**CASE STUDY**

**Predicting Where Crime Will Occur**

If you have seen the film *Minority Report*, you have gotten a far-fetched glimpse of a world where people are arrested for criminal acts that they are predicted to do, not that they have actually done. FOX also had a television series called *Minority Report* that was based on the same premise. While crime predictions in these shows are based on clairvoyants (people who can see into the future) and not real data, law enforcement agencies are beginning to use big data to predict both future behavior in individuals and, as we saw with crime mapping, areas where crime is likely to occur in the future.

As we highlighted earlier, crime mapping allows law enforcement agencies to estimate where hot spots of crime are occurring—where they have been most likely to occur in the past. Caplan and Kennedy (2015) from the Rutgers School of Criminal Justice have pioneered...
a new way to forecast crime using big data called **risk-terrain modeling (RTM)**. Using data from several sources, this modeling predicts the probability of crime occurring in the future using the underlying factors of the environment that are associated with illegal behavior. The important difference between this and regular crime mapping is that it takes into account features of the area that enable criminal behavior.

The process weights these factors (which are the independent variables) and places them into a final model that produces a map of places where criminal behavior is most likely to occur. In this way, the predicted probability of future crime is the dependent variable. This modeling is essentially special-risk analysis in a more sophisticated form than the early maps of the Chicago school presented earlier. Kennedy and his colleagues (2012) explain:

Operationalizing the spatial influence of a crime factor tells a story, so to speak, about how that feature of the landscape affects behaviors and attracts or enables crime occurrence at places nearby to and far away from the feature itself. When certain motivated offenders interact with suitable targets, the risk of crime and victimization conceivably increases. But, when motivated offenders interact with suitable targets at certain places, the risk of criminal victimization is even higher. Similarly, when certain motivated offenders interact with suitable targets at places that are not conducive to crime, the risk of victimization is lowered. (24)

Using data from many sources, RTM statistically computes the probability of particular kinds of criminal behavior occurring in a place. For example, Exhibit 11.8 displays a risk-terrain map that was produced for Irvington, New Jersey. From the map, you can see that several variables were included in the model predicting the potential for shootings to occur, including the presence of gangs and drugs, along with other infrastructure information such as the location of bars and liquor stores. Why were these some of the factors used? Because previous research and police data indicated that shootings were more likely to occur where gangs, drugs, and these businesses were present. This does not mean that a shooting will occur in the high-risk areas, it only means that it is more likely to occur in these areas compared to other areas. RTM is considered a use of big data because it examines multiple datasets that share geographic location as a common denominator.

**CASE STUDY**

**Predicting Recidivism With Big Data**

As you learned in Chapter 2, the Minneapolis Domestic Violence Experiment and the National Institute of Justice’s Spousal Abuse Replication Project, which were experiments to determine the efficacy of different approaches to reducing recidivism for intimate partner violence (IPV), changed the way IPV was handled by law enforcement agencies across the United States and across the globe. No longer were parties simply separated at the scene, but mandatory arrest policies were implemented in many jurisdictions across the country, which “swamped the system with domestic violence cases” (Williams and Houghton 2004, 438). In fact, some states now see thousands of perpetrators arrested annually for assaults against their intimate partners. Many jurisdictions are attempting to more objectively determine whether these perpetrators present a risk of future violence should they be released on parole or probation.

Williams (2012) developed one instrument to determine this risk, which is called the **Revised Domestic Violence Screening Instrument (DVSI-R)**. To determine the effectiveness of the DVSI-R in predicting recidivism, Stansfield and Williams (2014) used a huge dataset that would be deemed *big data*, since it contains information on 29,317 perpetrators arrested on family violence charges in 2010 in Connecticut and is continuously updated for recent arrests.
and convictions. To measure the risk of recidivism for new family violence offenses (NFVO), the DVSI-R includes eleven items: Seven measure the behavioral history of the perpetrator, including such things as prior nonfamily assaults, arrests, or criminal convictions; prior family violence assaults, threats, or arrests; prior violations of protection orders; the frequency of family violence in the previous six months; and the escalation of family violence in the past six months. The other four items include substance abuse, weapons or objects used as weapons, children present during violent incidents, and employment status. The range of the DVSI-R is from 0 to 28, with 28 representing the highest risk score.

To examine how the DVSI-R predicted future arrests, Stansfield and Williams (2014) used two measures of recidivism during an 18-month follow-up: rearrests for NFVOs and rearrests for violations of protective or restraining orders only. Results indicated that of the over 29,000 cases, nearly one in four (23%) perpetrators were rearrested, with 14% of those rearrested for a violation of a protective order. Perpetrators who had higher DVSI-R risk scores were more likely to be rearrested compared to those with lower risk scores. As you can see, using big data to improve decision making by criminal justice professionals is not a thing of the future, it is happening now. The availability of big data and advanced computer technologies for its analysis mean that researchers can apply standard research methods in exciting new ways, and this trend will only continue to grow.

**Exhibit 11.8  Risk-Terrain Map**

*Source: Obtained from personal correspondence with Leslie Kennedy and Joel Caplan.*
Subject confidentiality is a key concern when original records are analyzed with either secondary data or big data. Whenever possible, all information that could identify individuals should be removed from the records to be analyzed so that no link is possible to the identities of living subjects or the living descendants of subjects (Huston and Naylor 1996, 1698). When you use data that have already been archived, you need to find out what procedures were used to preserve subject confidentiality. The work required to ensure subject confidentiality probably will have been done for you by the data archivist. For example, the Inter-University Consortium for Political and Social Research (ICPSR) examines all data deposited in the archive carefully for the possibility of disclosure risk. All data that might be used to identify respondents are altered to ensure confidentiality, including removal of information such as birth dates or service dates, specific incomes, or place of residence that could be used to identify subjects indirectly (see http://www.icpsr.umich.edu/icpsrweb/content/ICPSR/access/restricted/index.html). If all information that could be used in any way to identify respondents cannot be removed from a data set without diminishing data set quality (e.g., by preventing links to other essential data records), ICPSR restricts access to the data and requires that investigators agree to conditions of use that preserve subject confidentiality. Those who violate confidentiality may be subject to a scientific misconduct investigation by their home institution at the request of ICPSR (Johnson and Bullock 2009, 218). The UK Data Archive provides more information about confidentiality and other human subjects protection issues at its website (https://www.ukdataservice.ac.uk/manage-data/legal-ethical).

It is not up to you to decide whether there are any issues of concern regarding human subjects when you acquire a dataset for secondary analysis from a responsible source. The Institutional Review Board (IRB) for the protection of human subjects at your college or university or other institution has the responsibility to decide whether they need to review and approve proposals for secondary data analysis. Data quality is always a concern with secondary data, even when the data are collected by an official government agency, and even when the data are “big.” Researchers who rely on secondary data inevitably make trade-offs between their ability to use a particular dataset and the specific hypotheses they can test. If a concept that is critical to a hypothesis was not measured adequately in a secondary data source, the study might have to be abandoned until a more adequate source of data can be found. Alternatively, hypotheses or even the research question itself may be modified to match the analytic possibilities presented by the available data (Riedel 2000, 53). For instance, digital data may be unrepresentative of the general population due to socioeconomic differences between those who use smartphones or connect to the Internet in other ways and those who live offline, as well as because of the difference between datasets to which we are allowed access and those that are controlled by private companies (Lewis 2015). Social behavior online may also not reflect behavior in the everyday world.

Political concerns intersect with ethical practice in secondary data analyses. How are race and ethnicity coded in the U.S. Census? You learned in Chapter 4 that changing conceptualizations of race have affected what questions are asked in the census to measure race. This data collection process reflects, in part, the influence of political interest groups, and it means that analysts using the census data must understand why the proportion of individuals choosing other as their race and the proportion in a multiracial category has changed. The same types of issues influence census and other government statistics collected in other countries.

Big data also creates some new concerns about research ethics. When enormous amounts of data are available for analysis, the procedures for making data anonymous no longer ensure that it stays that way. For example, in 2006, AOL released (for research purposes) 20 million search queries from 657,000 users after all personal information had been erased and only a unique numeric identifier remained to link searches. However, staff of The New York Times conducted analyses of sets of search queries and were able to quickly identify a specific
individual user by name and location, based on the searches. The collection of big data also makes surveillance and prediction of behavior on a large scale possible. Crime control efforts and screening for terrorists now often involve developing predictions from patterns identified in big data. Without strict rules and close monitoring, potential invasions of privacy and unwarranted suspicions are enormous (Mayer-Schönberger and Cukier 2013).

CONCLUSION

In our data-driven world, data generally—and the methods examined in this chapter specifically—are increasingly being used in research and in intelligence-led policing. For example, using SNA and crime mapping are now common techniques used in large police agencies to not only respond to crime but to also reduce, disrupt, and prevent it. As one police investigator explained,

Any data that we can collate online, whether it be that online evidence that may indicate the commission of offense or assist in making a nexus, a link to that offense, such as photographs, emails, text messages... We use any data that we can get our hands on lawfully, certainly, to assist in our investigations. (Chan and Moses 2017, 305)

The use of such methodological techniques is requiring police academies to incorporate data analysis components into their training. So even if you do not plan to become a researcher yourself, you will likely be required to make sense of data in virtually any career you pursue. Hopefully, you should now have the knowledge and skills required to find and use secondary data and to review analyses of big data to answer applied criminal justice and criminological research questions.

KEY TERMS

- Adjacency matrix 329
- Betweenness centrality score 333
- Big data 340
- Binary network 329
- Crime mapping 334
- Geographic information system (GIS) 334
- Intelligence-led policing 338
- Ngrams 341
- Nodelist 329
- Nodes 329
- Relational data 329
- Relations 329
- Risk-terrain modeling (RTM) 342
- Social network analysis 328
- Social networks 328
- Sociogram 329

HIGHLIGHTS

- SNA uses relational data to examine the patterns in social relationships that individuals and groups have with each other.
- Crime mapping for research purposes is generally used to identify the spatial distribution of crime, along with the social indicators (such as poverty and social disorganization) that are similarly distributed across areas (e.g., neighborhoods, census tracts).
- Using huge datasets, often termed big data, to determine trends and patterns of social phenomena is becoming increasingly possible because of advanced computer technology. Big data and advanced computer technology have helped to advance crime mapping techniques in creative ways, including new predictive modeling techniques such as RTM.
EXERCISES

Test your understanding of the chapter content. Take the practice quiz.

1. Read the original article for one of the studies described in this chapter. Critique the article, using the article review questions presented in Appendix B as your guide. Focus particular attention on procedures for measurement, sampling, and establishing causal relations.

2. What are the similarities and differences between secondary data analysis and big data analysis? Do you feel one of these approaches is more likely to yield valid conclusions? Explain your answer.

3. As it was described in this chapter, SNA was used in assessing the investigation of the Green River killer. Can you think of other situations in which this type of analysis could help researchers or practitioners learn about or address a problem in society? Write these thoughts down in a one-page report.

4. Write down at least five different research questions that crime mapping can answer. Then, find an article that used crime mapping. Describe how GIS offered new and different information when compared to the other criminological research articles reviewed from previous chapter exercises.

5. In this chapter, you learned that SNA is a fairly new methodological technique that can account for relationships that both individuals and groups have with each other. Think about different friends you have now who have specific attitudes toward alcohol use. One person may not use alcohol at all, while another may drink alcohol frequently. Construct a graph depicting the past four weekends and the friends you hung out with for each weekend. Does your alcohol consumption change depending on the friends with which you are socializing? What are some ways in which a SNA could help explain the individual and group behavior differences and similarities when it comes to alcohol consumption?

6. Imagine that you are responsible for training police officers in the use of data for intelligence-led policing. In this training, you need to communicate at least five ways in which using data for intelligence-led policing could potentially lead to better or improved policing. Write down these five things and at least three challenges you believe you may encounter during this training on data-informed policing.

DEVELOPING A RESEARCH PROPOSAL

Add a mapping dimension to your proposed study.

1. For your research question, there is likely an opportunity to describe some phenomenon (e.g., crime rates) globally, nationally, or locally. Select the units of analysis you wish to describe (e.g., census areas, cities, states, or nations).

2. Review the possible sources of data for your project. Search the web and relevant government, historical, and international organization sites or publications. Search the social science literature for similar studies and read about the data sources that they used.

3. Your university or library may have some mapping software that is user friendly. For example, many libraries have a program called “PolicyMap” that already contains data in many domains, including crime. Users can request variables to be mapped using various units of analysis, including states, counties, and so on. Using software similar to this, attempt to create several maps using variables related to your research project.

WEB EXERCISES

1. Select a current topic and write a research question about this topic that could be answered with counts of words in books. Use the Google Ngrams program described in this chapter to answer your question. Discuss the limitations of your approach, including the words you searched for and the way in which you identified relationships.

2. The National Institute of Justice (http://www.ncjrs.gov) has a wealth of information on crime mapping.
Search the site for information on crime mapping and you will find that the site contains a multitude of information, including the latest technological advances in crime mapping strategies for police departments as well as full-text articles discussing recent research that uses crime mapping techniques. Select a report and summarize its findings.

3. The U.S. Census Bureau’s home page (http://www.census.gov) contains extensive reports of census data, including population data, economic indicators, and other information acquired through the U.S. Census. This website allows you to collect information on numerous subjects and topics that can be used to make comparisons among different states or cities. Find the QuickFacts option and choose your own state. Now pick the county in which you live and copy down several statistics of interest. Repeat this process for other counties in your state. Use the data you have collected to compare your county with other counties in the state. Write a one-page report summarizing your findings.

4. Go to the International Association of Law Enforcement Intelligence Analysts (IALEIA) website (http://www.ialeia.org). On the site, you can download papers and information regarding numerous topics, including crime mapping. Provide a review of at least one of these applications. You can also link to IALEIA’s journal to read the latest publications in this area.

5. Go to the Rutgers Center on Public Security (http://www.rutgerscps.org/) and then go to the RTM menu. There you will find several places to browse, including a “Topics” section that will direct you to recent journal articles that have used RTM. Find a recent publication and describe the methodology. What were the independent variables, and what was used as the dependent variable? Are there any other variables that could have been added to predict the dependent variable?

ETHICS EXERCISES

1. Big data begin as little data; that is, as the records of phone calls, Twitter posts, or pictures taken by individuals in their daily lives. What limitations should be imposed regarding access to and use of such data once they have become aggregated into massive datasets? Is removing explicit identifiers sufficient protection? When does access to big data violate rights to privacy?

2. In January 2012, Facebook conducted an experiment in which emotional cues were manipulated for 689,003 users. Some saw news stories and photos on Facebook’s homepage containing many positive words, while others saw negative, unpleasant words. The messages sent subsequently by these users were a little more likely to reflect the emotional tone of the words they had been chosen randomly to see. When this experiment was reported in the Proceedings of the National Academy of Sciences (Kramer, Guillory, and Hancock 2014), some people were outraged. What do you think of the ethics of this type of big data experiment?

SPSS OR EXCEL EXERCISES

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012 states data.sav</td>
<td>This state-level dataset compiles official statistics from various official sources, such as the census, health department records, and police departments. It includes basic demographic data, crime rates, and incidence rates for various illnesses and infant mortality for entire states.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>region9</td>
<td>The regional location of each state within the nine designated regions of the U.S.</td>
</tr>
<tr>
<td>perfampoverty</td>
<td>The percentage of families below the poverty line in a state</td>
</tr>
<tr>
<td>murderRt</td>
<td>The murder rate reported per 100,000 for a state</td>
</tr>
</tbody>
</table>
1. Dr. Smartypants says, “Let’s examine whether poverty rates vary by region!” You can examine the means for perfampoverty variable by using the means procedure. This is done by analyze->compare means->means and putting perfampoverty in the Dependent List box and region9 in the Layer 1 of 1 box.

   a. You can also visualize the regional differences in poverty by examining a bar graph. Do this by clicking on graphs->legacy dialogs->bar. Click on “Simple” and check “summaries for groups of cases” then define. Place perfampoverty in the Bars Represent box and check “other statistic” and the default will be the mean. Place region9 in the Category Axis box. Click OK, and hang on to this graph.

2. Now repeat these two procedures for the murder rates and region.

3. “Wow,” Dr. Smartypants thinks, “it would be cool to see if regions with high rates of poverty also have high rates of murder.” To visualize this, we can ask for a bar chart that depicts both the murder rate and percentage of families living below poverty by region together. To do this click graphs->legacy dialogs->bar. Click on “clustered” and check “summaries for separate variables” then define. Place both murderRt and perfampoverty in the Bars Represent box (the default will be to display the means) and place region9 in the Category Axis box. Click OK. Interesting. What would you conclude?